# Chasing strain variation (and what to do when you catch it)

C. Titus Brown

MBL STAMPS 2022

July 28th

# Warning! Forward-looking opinions ahead!

Prediction is very difficult, especially
if it's about the future!

- Niels Bohr


https://www.amazon.com/SmartSign-Caution-Falling-Cows-Funny-Road/dp/B087JRRR4V

# I’m really enthusiastic about these approaches!!!!

- But I/we won’t know for some time if the stuff I’ll show you is a good direction.
- That’s ok!

- Brainstorming and constructive criticism is very welcome! As are questions!

# Summary:

- Strain variation (pangenomic variation/accessory elements) will be key to understanding a large component of microbiome function.
- Reference- and assembly-based approaches both **fail** in obvious ways when trying to discover and interpret strain variation.
- K-mer, gene, and assembly-graph based approaches are promising ways forward.

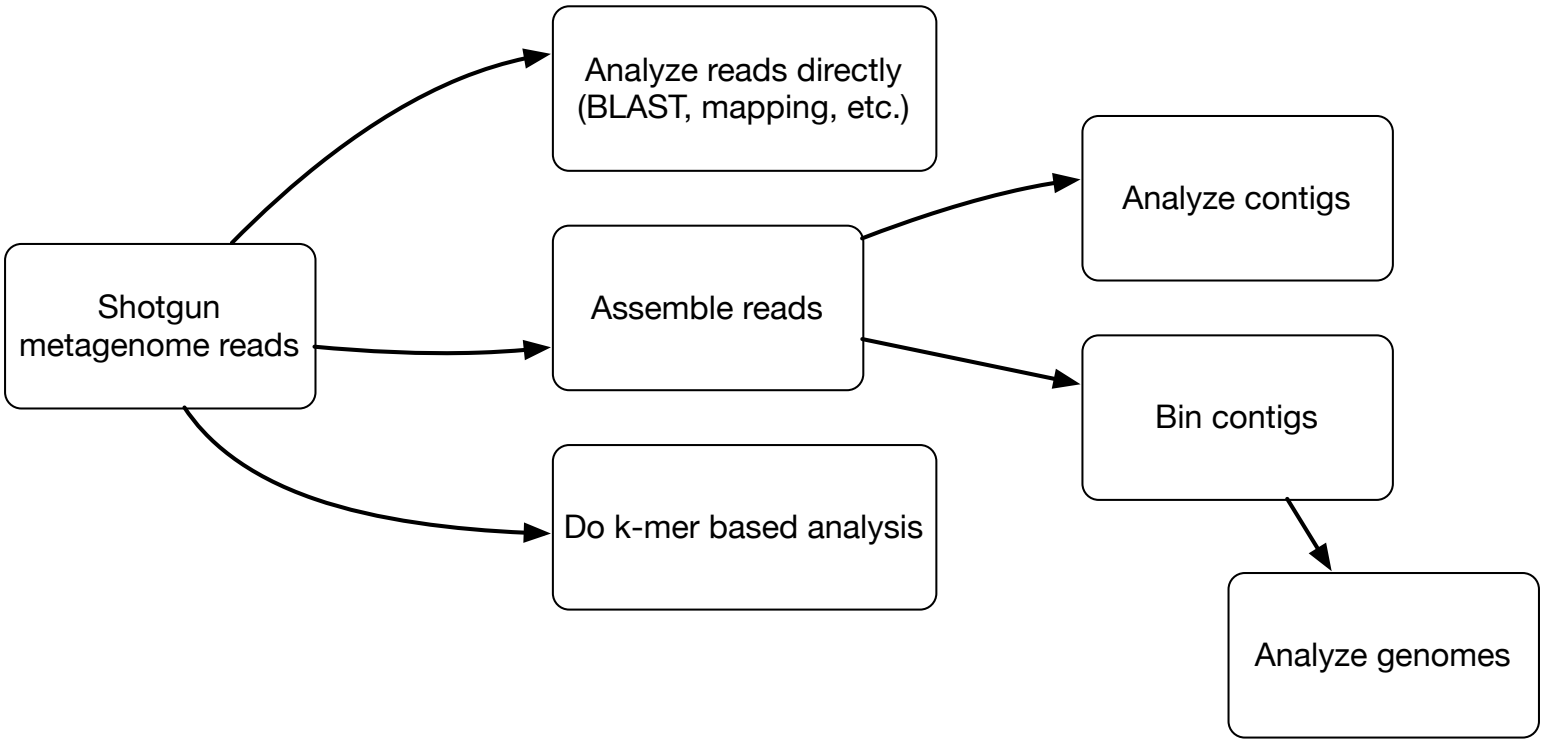
- …but we're still building the tools and ecosystem to support this. You should come join us in the k-mer cult! We're nice people!

# Two references that underlie *my* perspective -

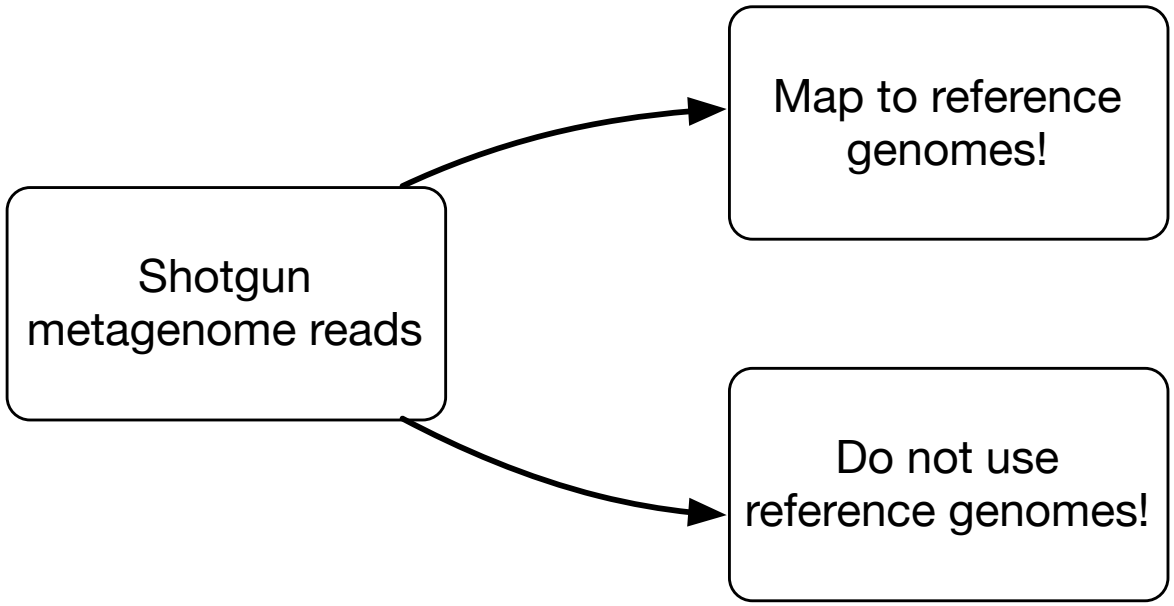
- "**Meta-analysis of metagenomes via machine learning and assembly graphs reveals strain switches in Crohn's disease"**
  - Reiter et al., 2022 (with Tessa and Amy :)
  - https://www.biorxiv.org/content/10.1101/2022.06.30.498290v1
- "**Protein k-mers enable assembly-free microbial metapangenomics"**
  - Reiter et al., 2022 (with Tessa)
  - https://www.biorxiv.org/content/10.1101/2022.06.27.497795v1

tl;dr: This perspective emerges from Taylor's unrelenting drive on her thesis work, which is partly encapsulated in the above papers.

# To assemble, or not to assemble?

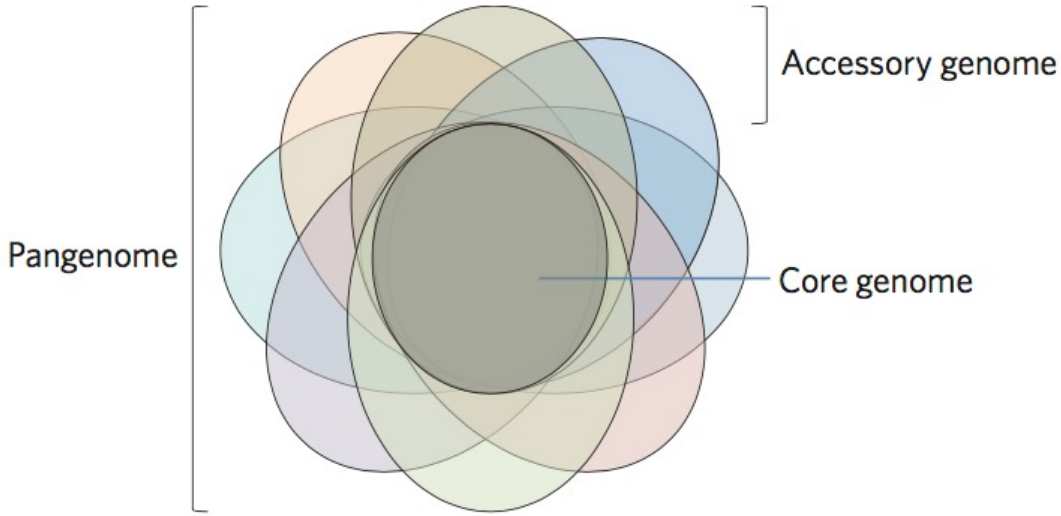


# Should you use reference genomes?

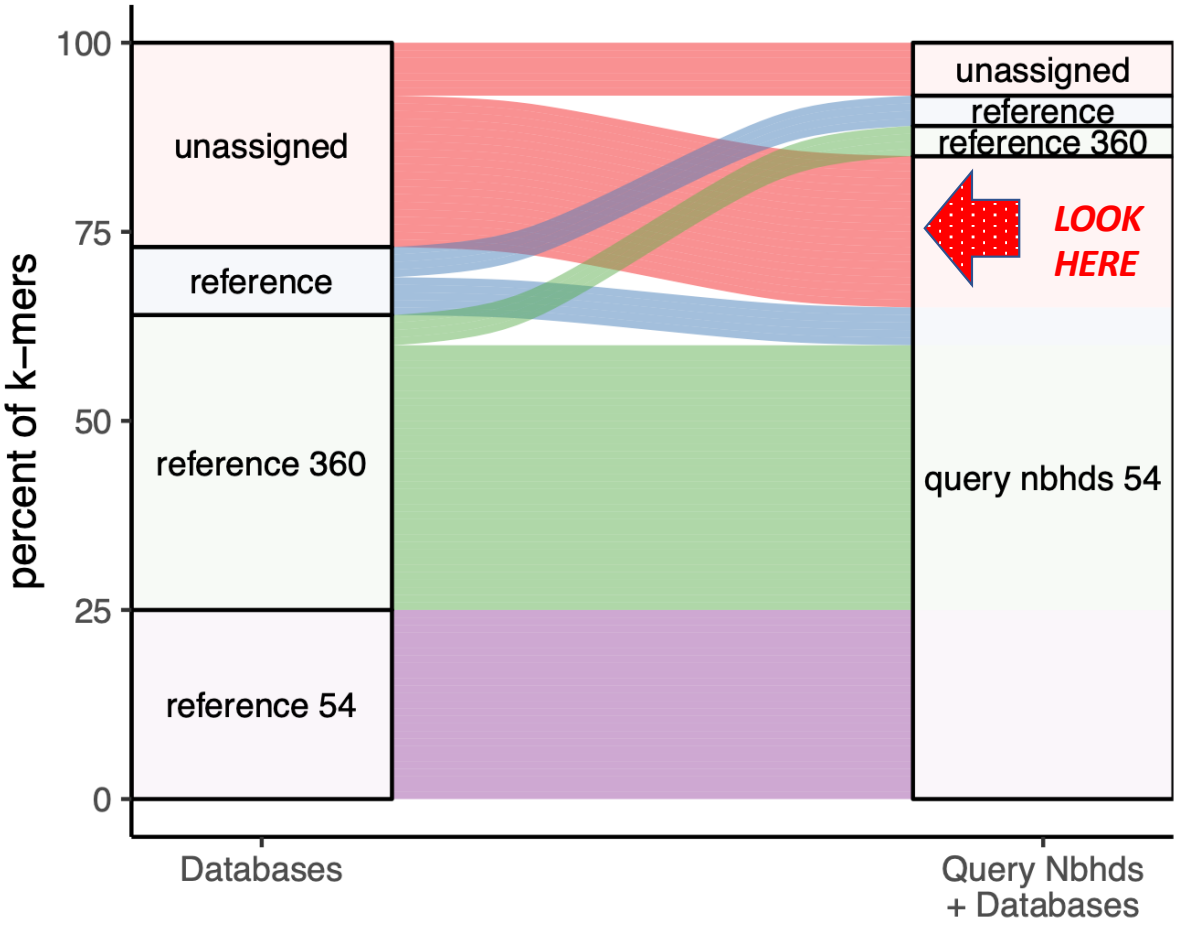


# Background: Pangenomes and metapangenomics

Most (all?) microbial species have significant amounts of strain variation.




McInerny et al., DOI: 10.1038/nmicrobiol.2017.40




From Reiter et al, 2022, 10.1101/2022.06.30.498290 (IBD paper)

**tl;dr**

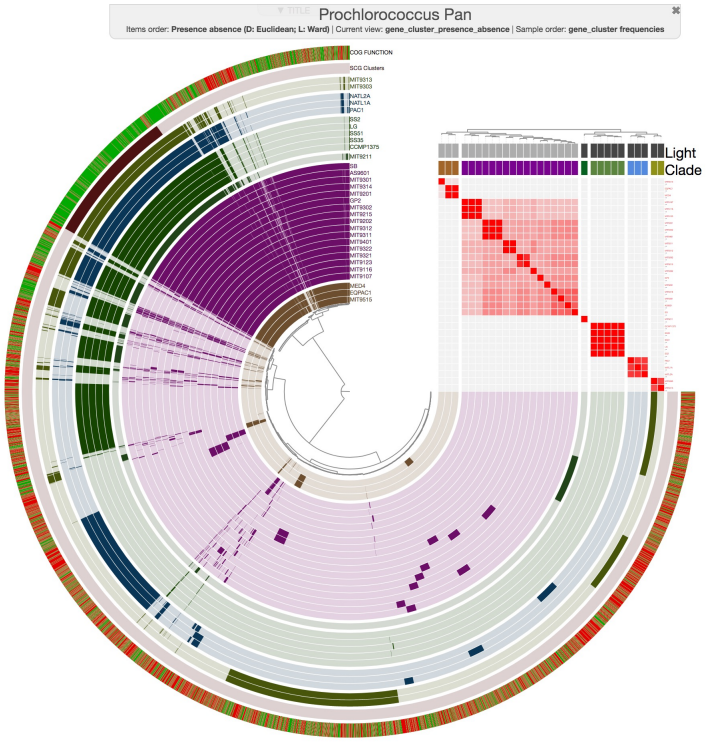
**About 20% of important k-mers are probably in unknown accessory elements**

Let’s explore this!

# Backing up a bit, here are a few observations -

- Any given reference genome is a snapshot of a microbial strain in a given time/place.
  - This is true of isolates, single-cell genomes, and MAGs!
  - There is no simple way to measure the accurate inclusion of non-core/single-copy genes in MAGs! So accessory element inclusion is ???
- Metagenomes, however, *do* (or at least *may*) contain the relevant accessory elements!
  - Microbiomes gotta do their thing, yo.
- But *assembling* them and *binning* them is tricky and not guaranteed.
  - Strain confusion challenges assembly…
  - Binning is largely evaluated only on inclusion of core genes/completeness.

# e.g. *Prochlorococcus* species/strain clades

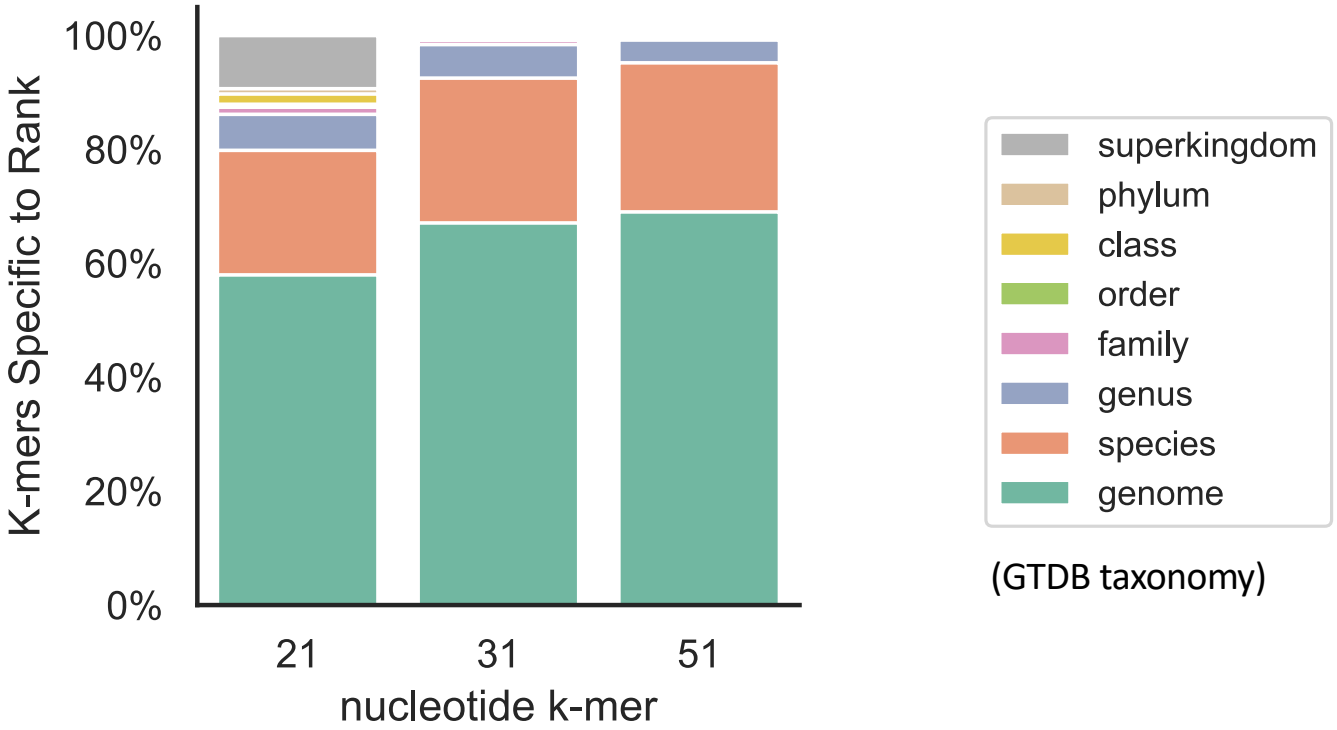



Clusters based on gene/presence absence in different *Prochlorococcus* genomes.

This *must* be happening in metagenomes!

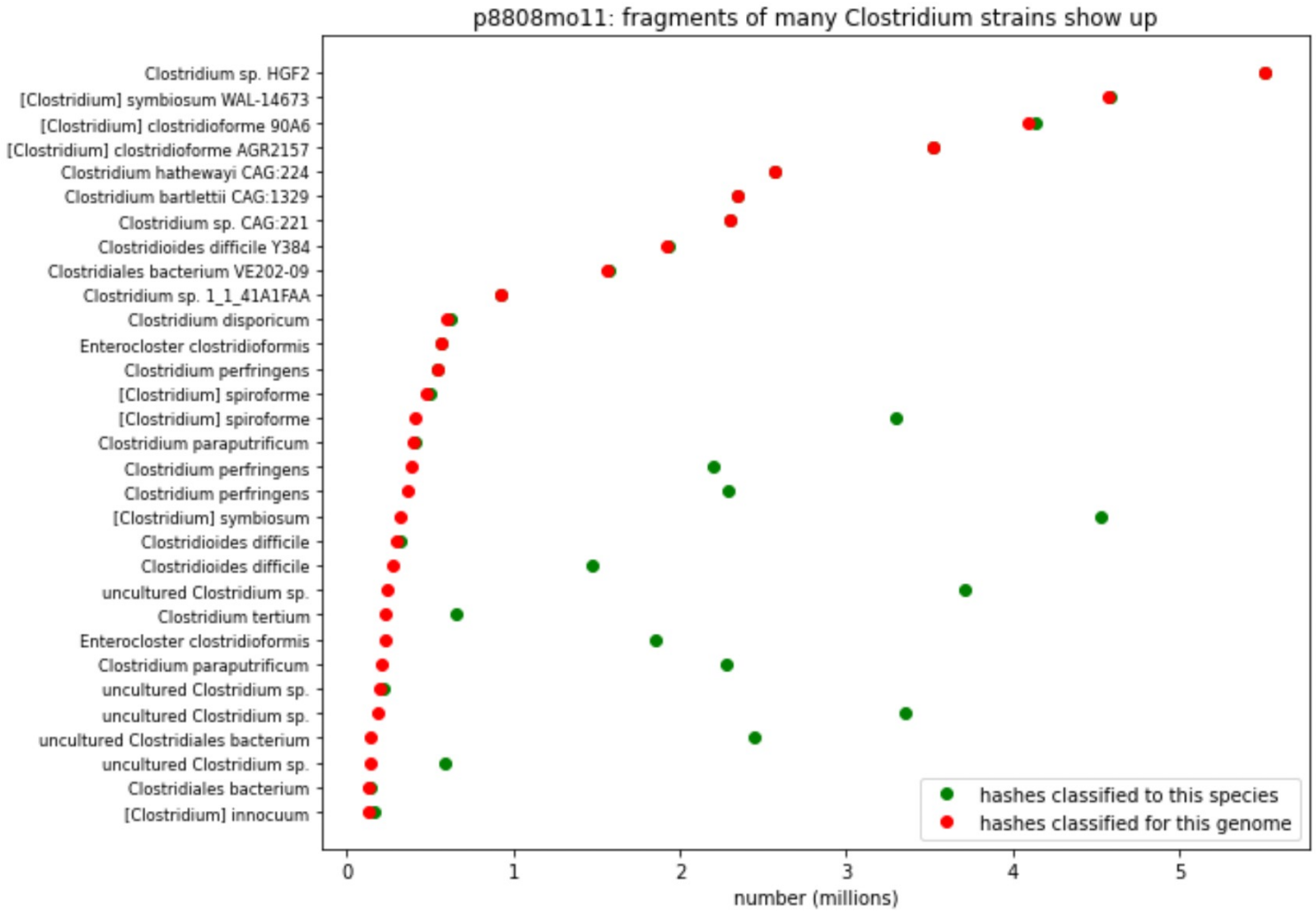
https://merenlab.org/2016/11/08/pangenomics-v2/

# Most of a genome's k-mers are specific to that genome…




(GTDB taxonomy)

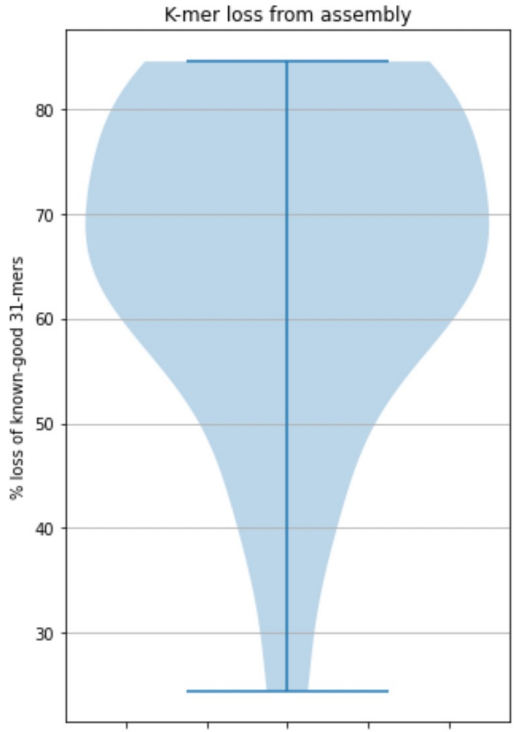
Pierce-Ward et al., in preparation.




Metagenomes contain “shards” of many different genomes from the same species.

These are pangenomic elements that are part of the strain(s) in the metagenome.

# ~30% of “good” 31-mers are lost during metagenome assembly




We see substantial loss of k-mers during *assembly*.

Our tentative interpretation is that this is due to a mixture of:

- Abundance filtering (by assembly)
- Collapse of strain variation

(We have not measured information loss during binning but it is ***guaranteed*** to be substantial.)

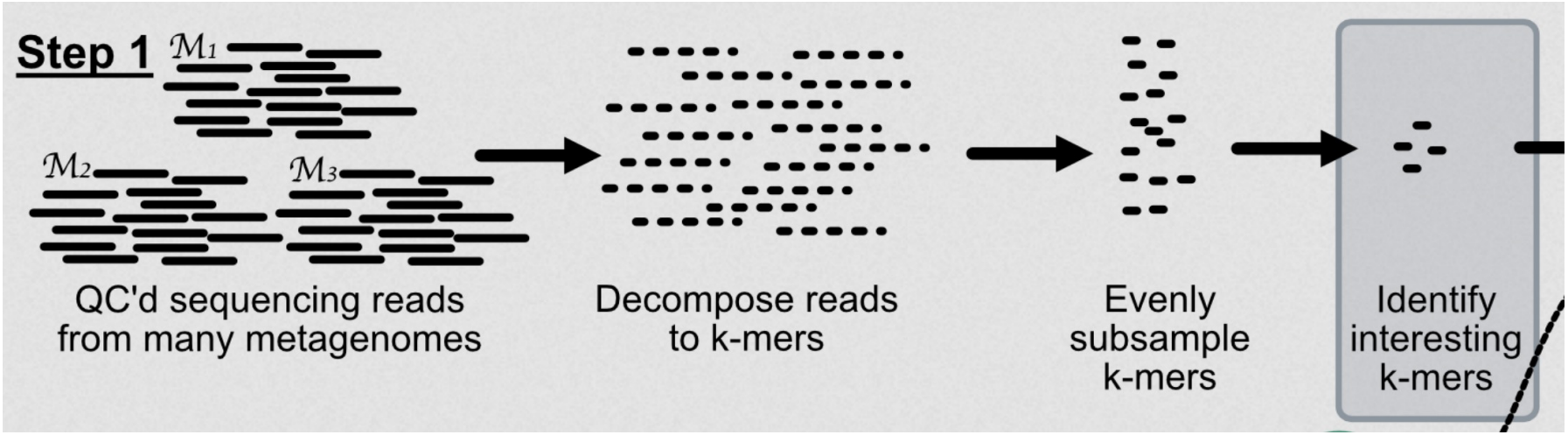
Unpublished dib-lab work.

# Interim conclusion: you can't rely on reference- or assembly-based approaches to study strains in metagenomes.

- Strain variation and accessory elements are important to examine in a wild-microbe context – **metagenomes.**
- You *have* to use assembly-free approaches – k-mers, or gene-based approaches, or ?? – to dig into strain variation in metagenomes.
- (Note that 16S do not, and marker genes *may not,* have the necessary resolution to detect strain variation. Regardless, they will not generally encode niche-specific function.)

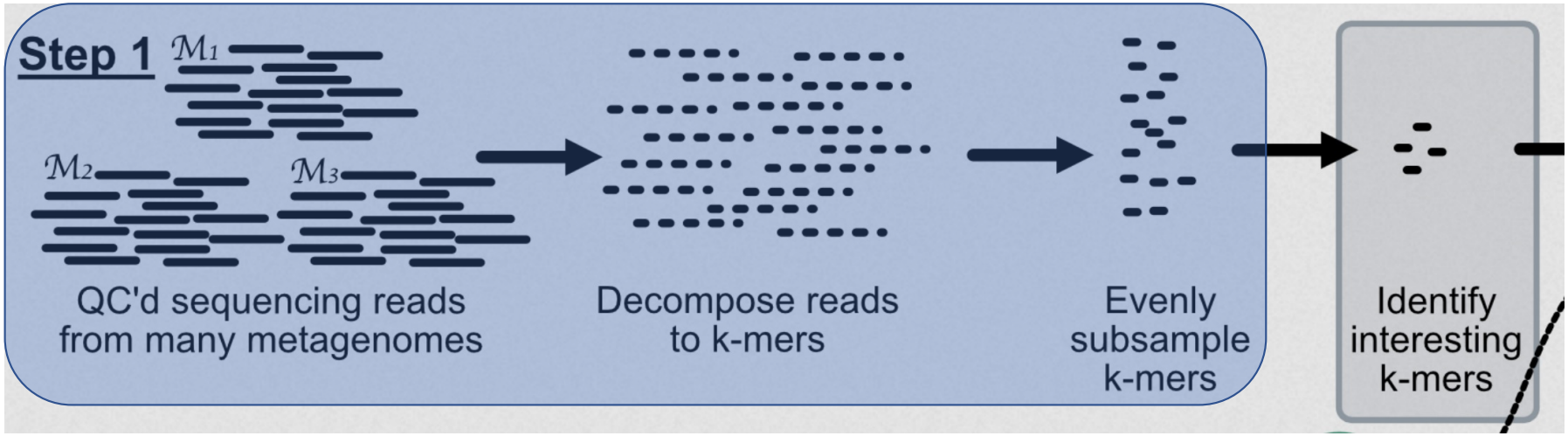
- So what do we suggest?

# One approach (Taylor's approach :)



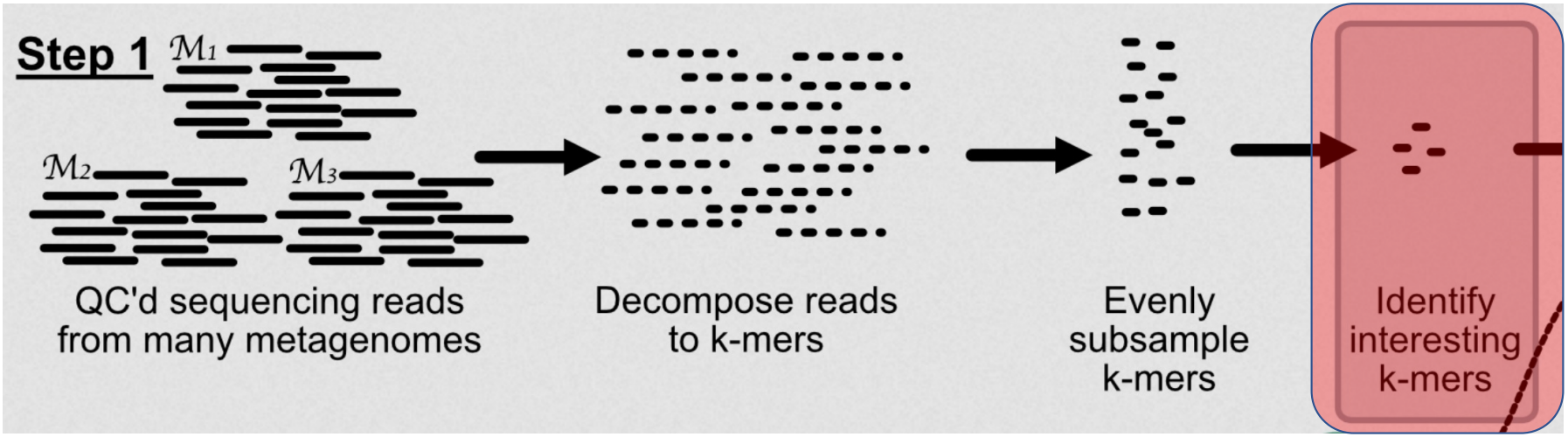

From Reiter et al, 2022, 10.1101/2022.06.30.498290 (IBD paper)

# One approach (Taylor's approach :)




You all know how to do this now – QC + sourmash!

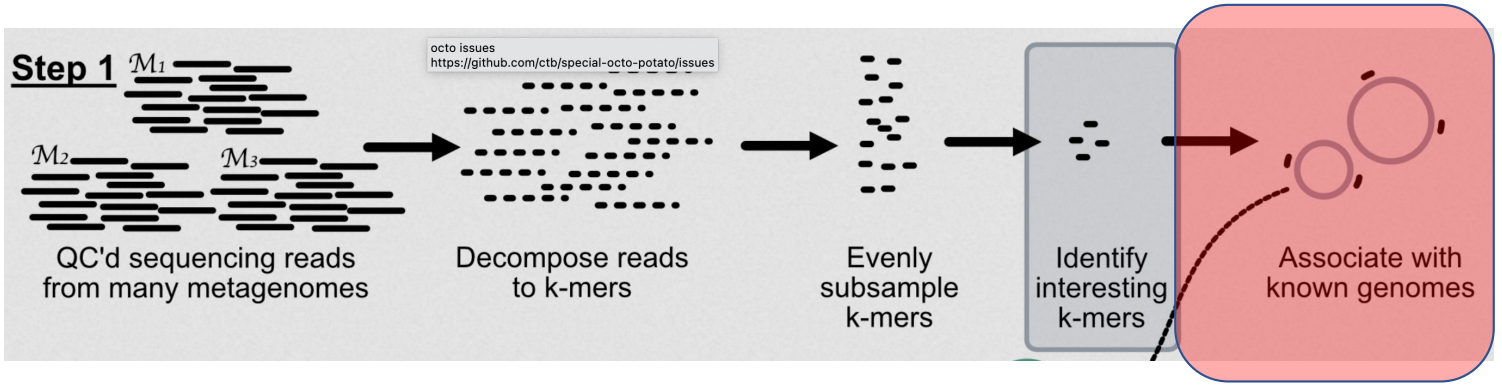
# One approach (Taylor's approach :)



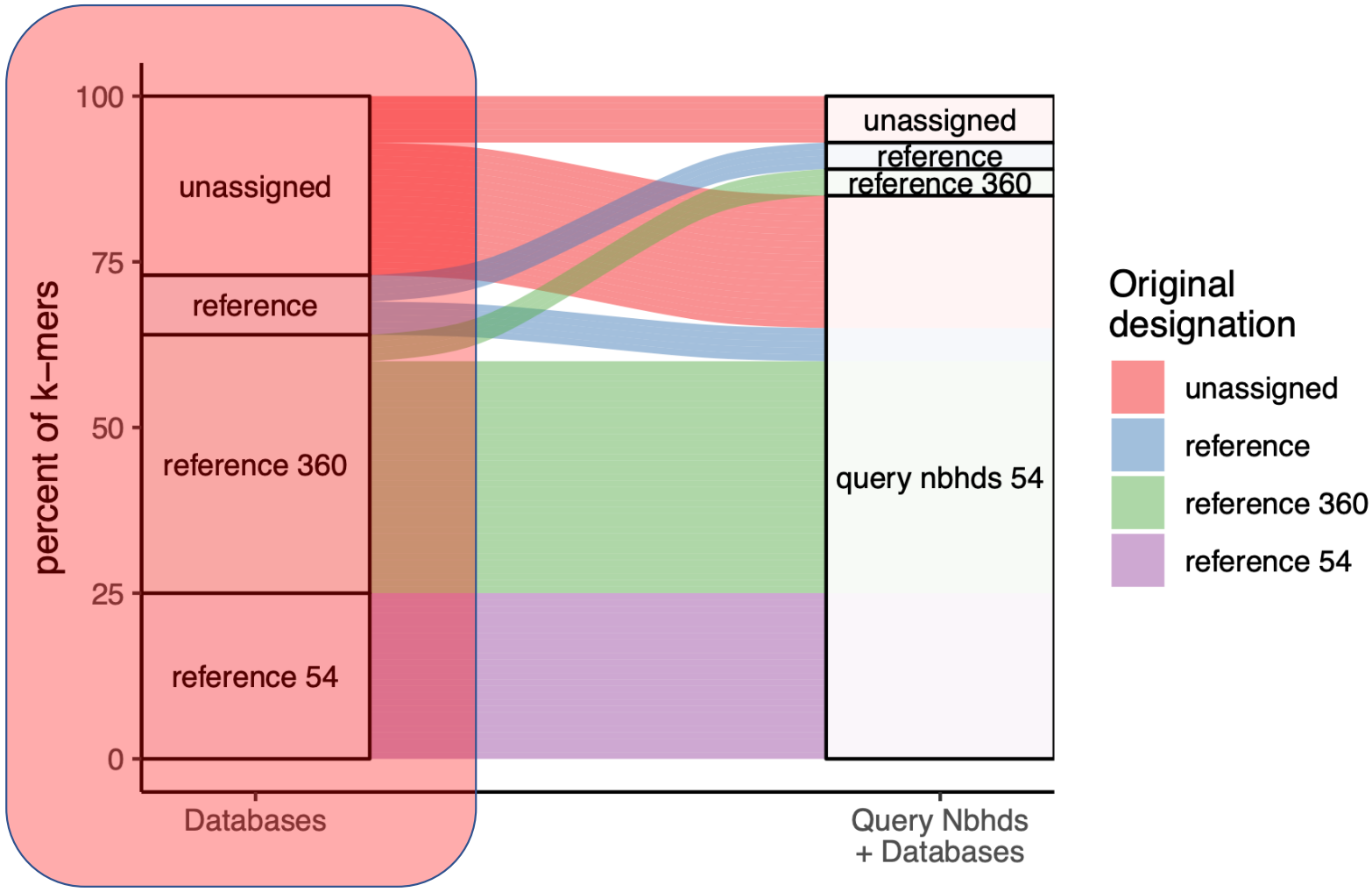

Taylor used random forests (machine learning) to do this.

# This leaves you in a bit of a pickle…

K-mers on their own aren't very informative… need to link to (something)!




Y'all know how to do this now, too – sourmash gather!




From Reiter et al, 2022, 10.1101/2022.06.30.498290 (IBD paper)

**We could only identify ~75% of interesting k-mers, even when using *all* the reference genomes.**

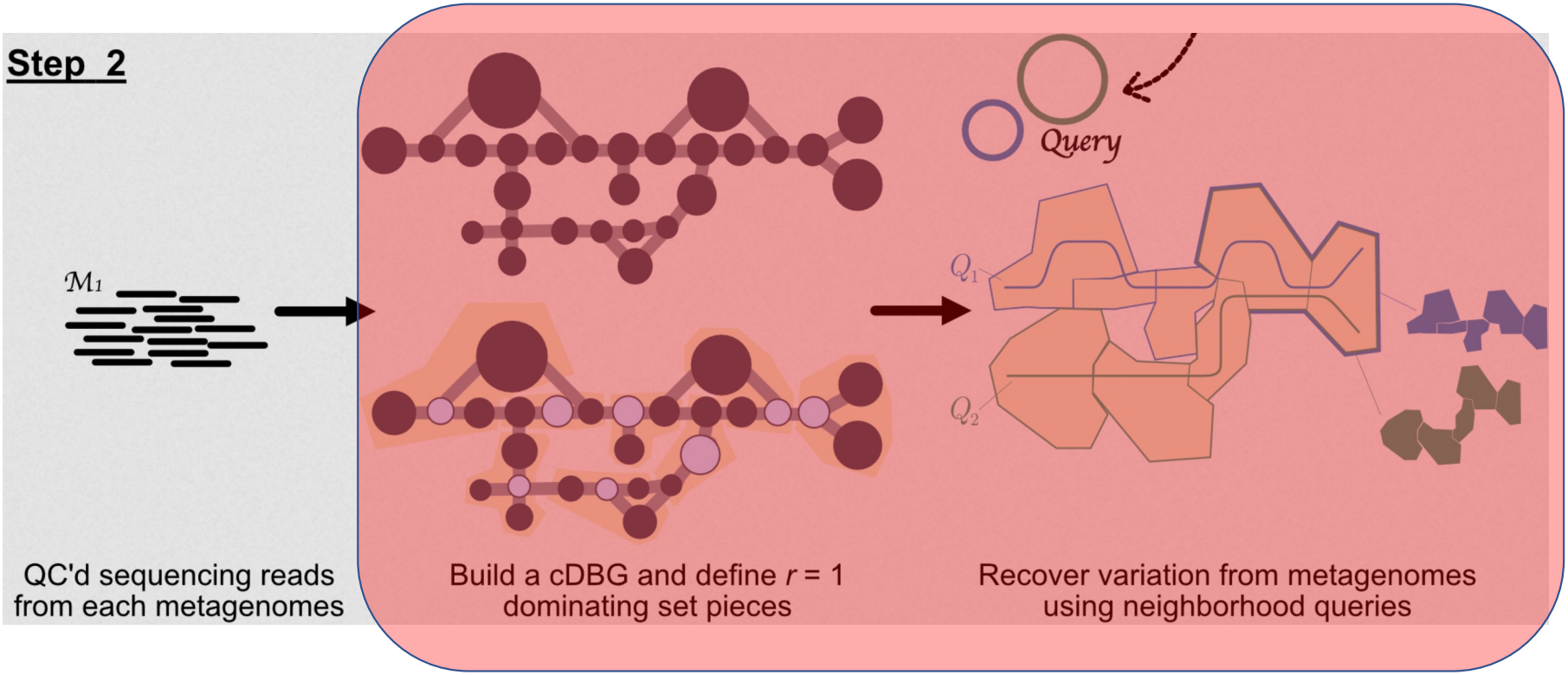
# How now, brown cow?

- Well this is still a pickle…
- …25% of “interesting” k-mers are unannotated!

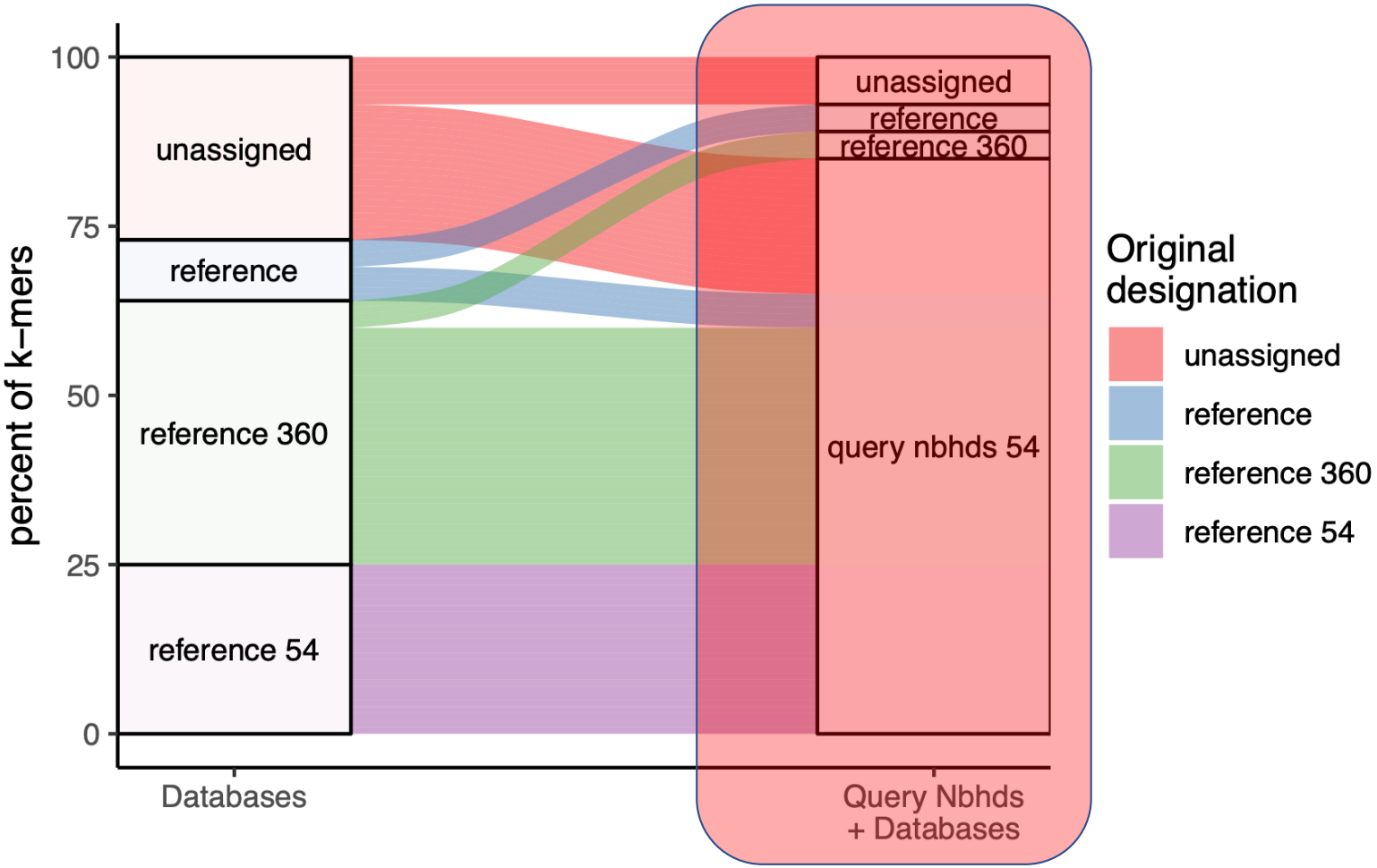
- Taylor tried *everything*. Seriously. Assembly into contigs; mapping; anger; despair; the Dark Side.

- The only thing that worked was using *assembly* ***graph*** *neighborhoods*.

# Spacegraphcats software enables *reference-guided* retrieval from assembly graphs.



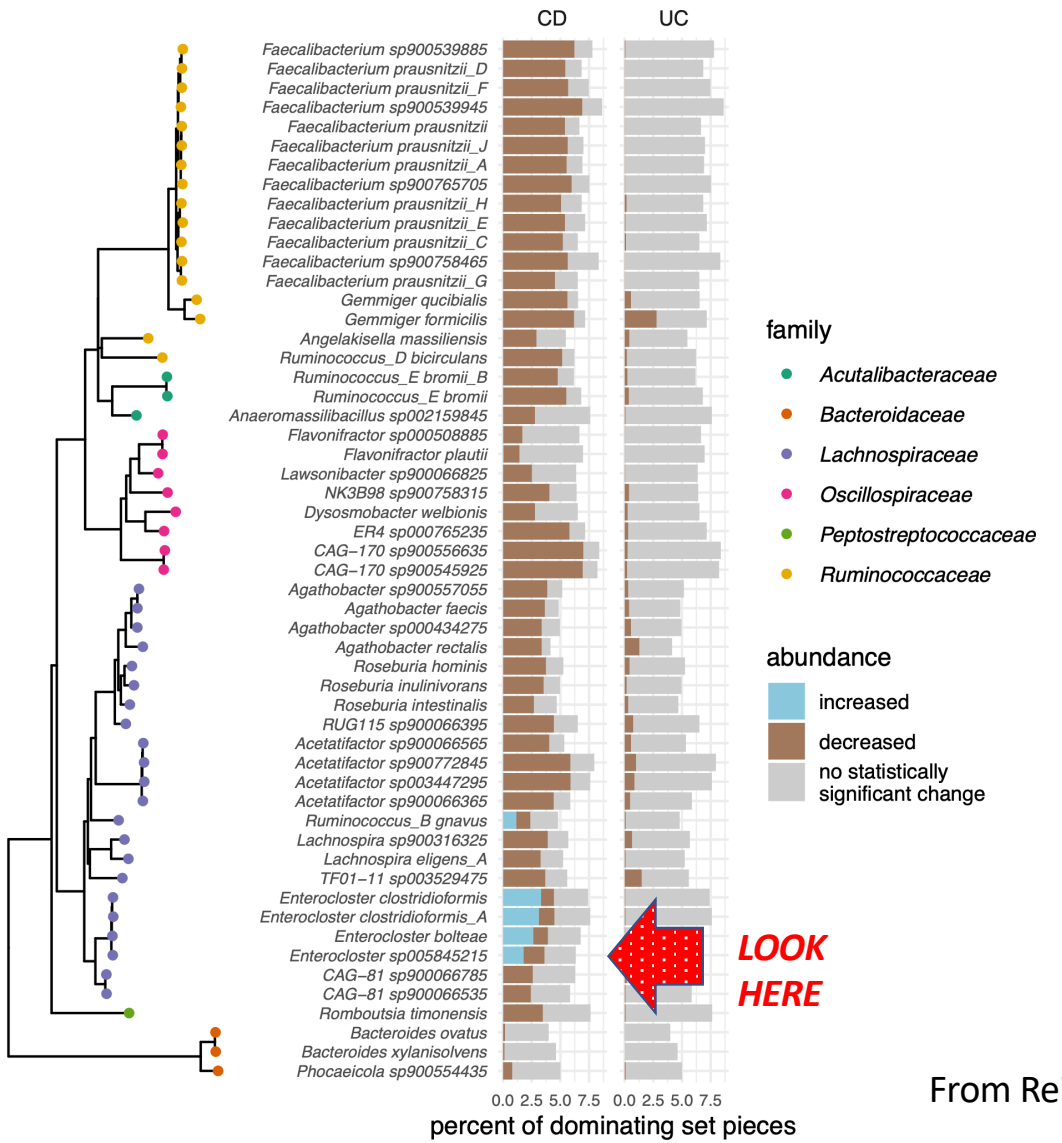

Brown et al., 2020: 10.1186/s13059-020-02066-4




From Reiter et al, 2022, 10.1101/2022.06.30.498290 (IBD paper)

**This let us associate another 15% of k-mers to genomes –**

**We *infer* that these k-mers belong to accessory elements.**



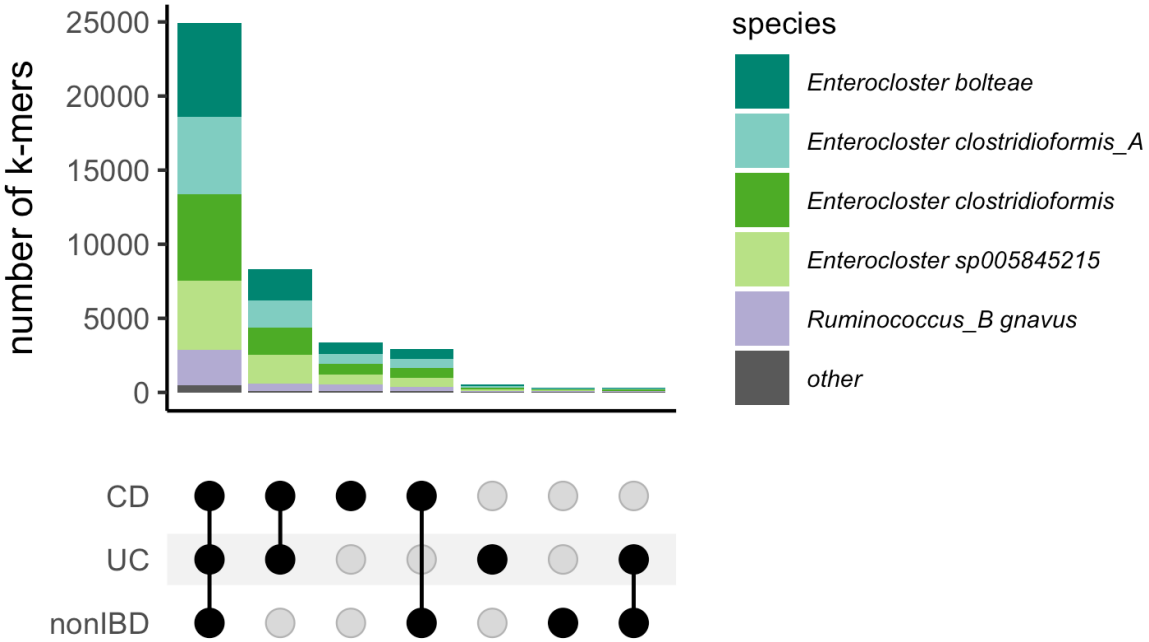

**In a *background* of overall decreasing strain abundance in Crohn's Disease metagenomes (vs non-IBD), we see statistically significant *increases* in the abundance of five strains.**

Our conclusion: this is *strain switching,* and these strain switches correlate with Crohn's Disease.

From Reiter et al, 2022, 10.1101/2022.06.30.498290 (IBD paper)

# (There's no magic bullet, unfortunately)

**Most differentially abundant sequences occur in metagenomes of individuals diagnosed with CD, UC and non-IBD.**



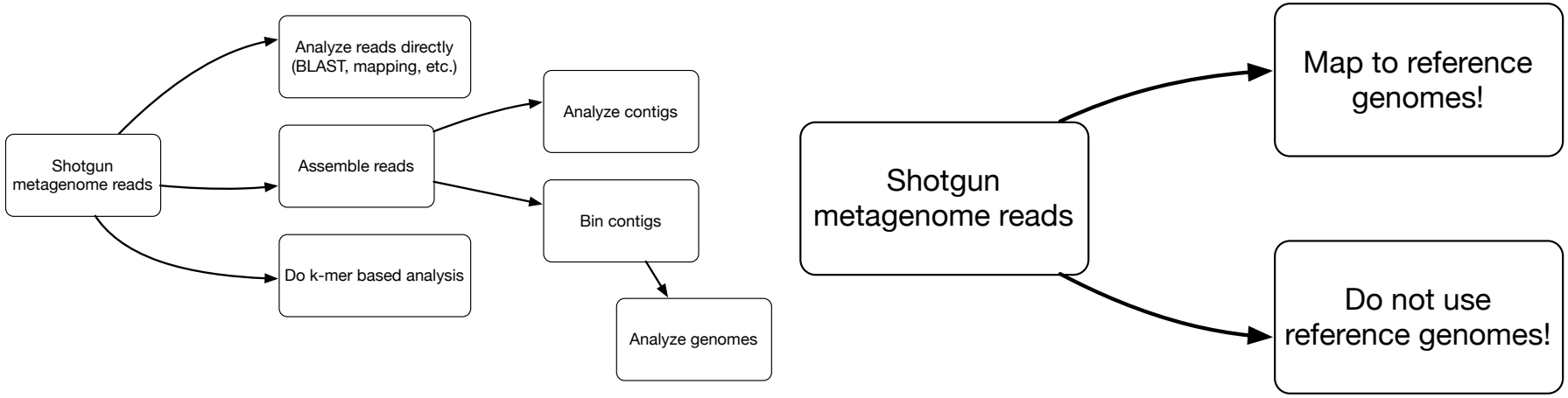

From Reiter et al, 2022, 10.1101/2022.06.30.498290 (IBD paper)

# Further interim conclusions

- Even in the human gut, there are many ~unknown sequences that (may) play an important role in microbiomes.
- Substantial portions of these sequences are not represented in current reference genome catalogs.
- Assembly isn't capable of robustly recovering these sequences.

- At the same time, let me just say that it's virtually impossible to interpret metagenome data without using reference genomes or assembly 😂😭

# These are false choices! You need to do it all!



# Our recommendation(s) for > 2022

- Characterize your data in reference-independent ways *first.*
- Find/analyze statistically significant patterns.
- Then *interpret* those patterns using reference genomes, assembly, etc as you need.
- This approach lets you quantify information loss due to various approaches!

- Unfortunately, tools and workflows to do this all robustly are not very mature…

# Tools to keep an eye on re graphs and metapangenomes -

- Sourmash and spacegraphcats
- Metagraph
- STRONG
- KOMB (Todd)
- MetaCortex
- metacherchant

# Taylor's documentation and tutorials

- **How to read/interpret a metagenome assembly graph:** Here is a primer on metagenomes and assembly graphs
- I want to build and analyze assembly graphs myself - https://spacegraphcats.github.io
- I want to run workflows that have been run before and see if they work for me https://github.com/spacegraphcats/2018-paper-spacegraphcats: metagenome bin completion (e.g. scooping in reads that didn't assemble)
- https://github.com/dib-lab/2021-metapangenome-example (metapangenome analysis with assembly graphs)
- https://github.com/dib-lab/2022-dominating-set-differential-abundance-example (differential abundance analysis on the graph)